Seminar on Distributed Data Management

Evica Ilieva and Sebastian Michel

Universität des Saarlandes
Sommersemester 2014

CONSULTANTS SAY THREE QUINTILLION BYTES OF DATA ARE CREATED EVERY DAY.

IT COMES FROM EVERYWHERE. IT KNOWS ALL.

ACCORDING TO THE BOOK OF WIKIPEDIA, ITS NAME IS “BIG DATA.”

BIG DATA LIVES IN THE CLOUD. IT KNOWS WHAT WE DO.
Big Data

- **Massive** amounts of data from a variety of sources
  - Web search logs
  - Social networks and blogs
  - RFID and other sensor data
  - Sales data
  - Scientific data
Big Data (Cont’d)

• Big data is often associated with NoSQL and MapReduce tools to process it.

• Processed in and across gigantic data centers.

• The term “Big Data” denotes not only size but things we want to/can do with it (benefits).
Huge Amounts of Data

• Google:
  – Billions of Websites
    (around 50 billion, Spring 2013)
  – TBs of data

• Twitter:
  – >100 million tweets per day

• Cern’s LHC
  – 25 Petabytes of data per year
Huge Amounts of Data(2)

- Megaupload
  - 28 PB of data
- AT&T (US Telecom. Provider)
  - 30 PB of data through its networks each day
- Facebook
  - 100 PB Hadoop cluster

http://en.wikipedia.org/wiki/Petabyte
Example: Reading 10TB from Disk

• Assume you got 10 TB data on disk
• Want to do some analysis of it

• With a 100MB/s disk, reading alone takes
  – 100000 seconds
  – 1666 minutes
  – 27 hours
Need to do something about it

http://flickr.com/photos/jurvetson/157722937/

http://www.google.com/about/datacenter
Scale-Out vs. Scale-Up

• Scale-Out (Many Servers -> Distributed)

• As opposed to Scale-Up
Scale-Out

• Common technique is scale-out
  – Many machines
  – Amazon’s EC2 cloud, around 400,000 machines

• Commodity machines (many but not individually super fast)

• Failures happen virtually at any time.

• Electricity is an issue (particularly for cooling)

http://huanliu.wordpress.com/2012/03/13/amazon-data-center-size/
Hardware Failures

• Lots of machines (commodity hardware) → failure is not exception but very common

• \( P[\text{machine fails today}] = \frac{1}{365} \)

• \( n \) machines: \( P[\text{failure of at least 1 machine}] = 1 - (1 - P[\text{machine fails today}])^n \)

  - for \( n=1 \): 0.0027
  - for \( n=10 \): 0.02706
  - for \( n=100 \): 0.239
  - for \( n=1000 \): 0.9356
  - for \( n=10000 \): ~ 1.0
New Requirements

• MapReduce is one prominent example that novel businesses have new requirements.

• Going away from traditional RDBMS.

• Addressing huge data volumes, processed in multiple, distributed (wide spread) data centers.
New Requirements (Cont’d)

• Massive amounts of unstructured (text) data
• Processed often in batches (with MapReduce).

• Huge graphs like Facebook’s friendship graph

• Often enough to store (key, value) pairs

• Massive streams of data (e.g., Twitter)
Wish List

• Data should always be consistent

• Provided service should be always quickly responding to requests

• Data can be (is) distributed across many machines (partitions)

• Even if some machines fail, the system should be up and running
Example:
CAP Theorem (Brewer's Theorem)

• System **cannot provide all 3 properties at the same time:**
  – Consistency
  – Availability
  – Partition Tolerance

Seminar Papers/Topics

- Cloud Computing
- Consistency Models
- MapReduce
- Continuous Data (Streams)
- Graph Data

11 very recent papers, capturing current trends and novel ideas
Seminar Goals, Regulations, ...
Goals

• Read, understand, and explore scientific literature
• Give presentations about your topic to your fellow students.
• Summarize a current research topic in a concise report
• Moderate a scientific discussion about a topic of one of your fellow students
• Receive feedback
Schedule

• 1 kick-off meeting on April 14th

• It is a block seminar, so:
  – talks will take place on two (consecutive) days
  – in early August
  – we will do a doodle poll soon to find those days
Organization and Requirements

• Attend the kick-off meeting. Register in HISPOS.

• Give a 45min talk about your topic.
  – Talk + slides in English.
  – Be able to answer questions (roughly 15 min)

• First meeting with tutor by end of May.
• It is your responsibility to schedule this.

• By end of June: Send complete slides to tutor
Organization and Requirements (2)

• **Moderate one talk:**
  – Introduce speaker
  – Lead discussion phase
  – Prepare a couple of own questions (as fallback)

• **Send slides to moderator** 2 weeks before seminar

• **Hand in** written report, not later than 3 weeks after the seminar.
Organization and Requirements (3)

• **Attend** the entire block seminar (**all talks!**) 

• **Actively participate** in the seminar
  – Ask questions
  – Give constructive feedback: How did you like the talk? Was it clear enough? Slides ok? Improvements?
  – Reading other papers helps (e.g., maybe 1-2 most related ones to your own paper)
First Meeting with Tutor

• No need to have full-fledged slides ready
• But:
  – have a solid understanding of the paper (not necessarily until the very last theorem or advanced technique)
  – bring outline of your talk
  – concrete questions if something is unclear
Talk

- See Evica’s talk
Report

• **Concise summary** of the paper.
  – Motivation/problem, key issues of approach, results. Add your *own conclusions* .... but no need to praise or kill it.

• **Not a copy** of the paper

• Not longer than **4 pages**. Latex template will be provided.
Final Grade Based on

• **Performance in talk**
  – Presentation style & slide quality
  – Knowledge about the topic
  – Question handling

• **Participation**
  – Performance as moderator
  – Participation in general

• **Written report**

<table>
<thead>
<tr>
<th>Component</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance in talk</td>
<td>60%</td>
</tr>
<tr>
<td>Participation</td>
<td>20%</td>
</tr>
<tr>
<td>Written report</td>
<td>20%</td>
</tr>
</tbody>
</table>
General Advice

• Don’t start late with the talk preparation!
• Make good use of your tutor
  – Don’t hesitate to ask for help/feedback.

• Tutor provides general guidance
• Is a partner for discussion
• Helps with the harder parts of the scientific article
• Provides feedback about your presentations + report.
• Ask your fellow students for feedback.
Good Scientific Practice

• Cite papers or books you use
• Clearly refer to sources of copied illustrations, quotes, etc.

• Create your own slides (although there might be slides available online)!

• In the talk and report, refer to this seminar at title slide/page